Zookeeper异地跨数据中心的架构选择

陈东明
饿了么北京研发中心

Agenda

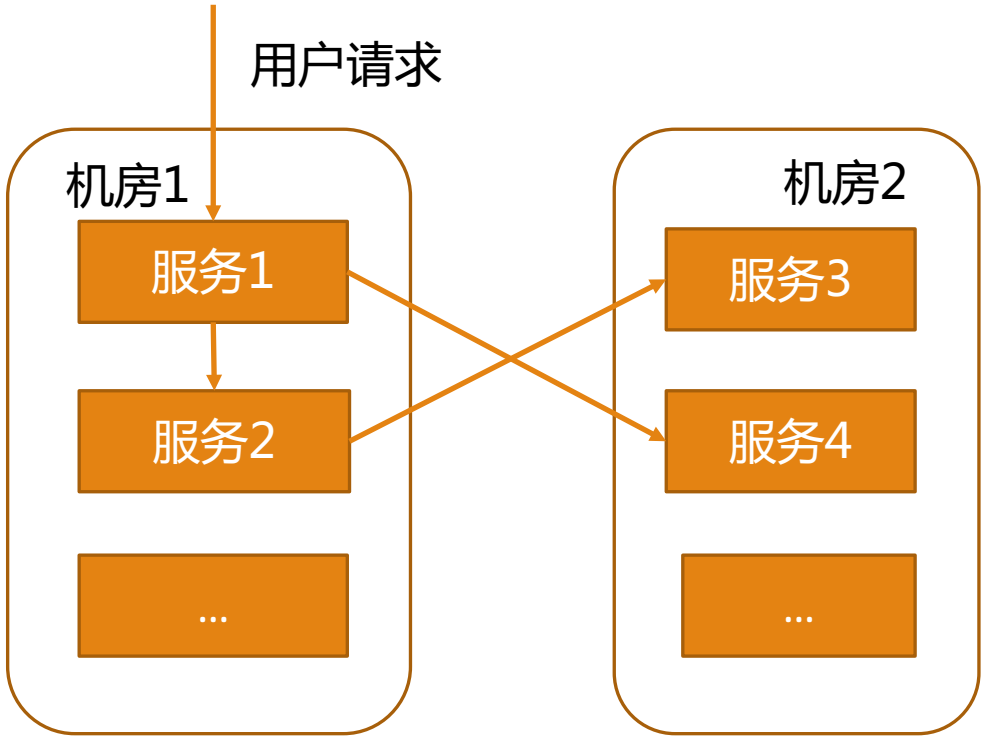
- 饿了么异地数据中心多活

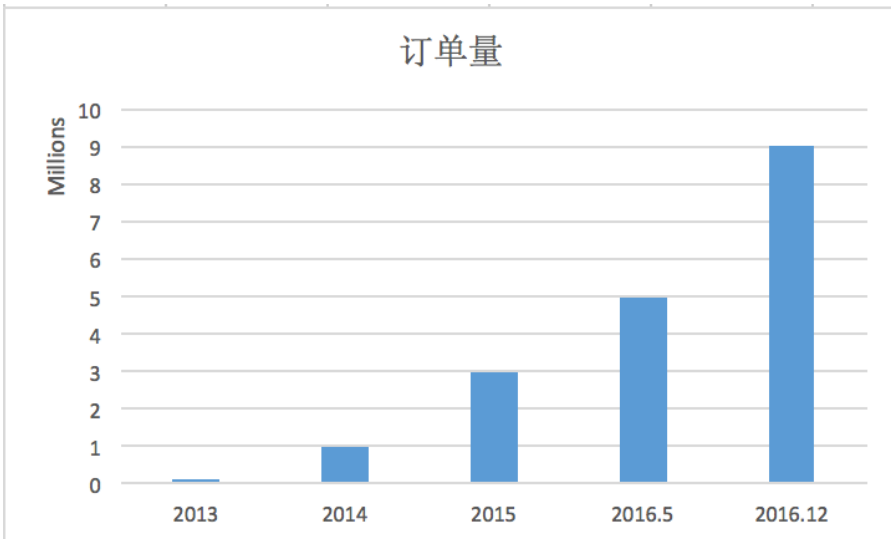
Zookeeper跨数据中心的方案

饿了么Zookeeper跨数据中心方案

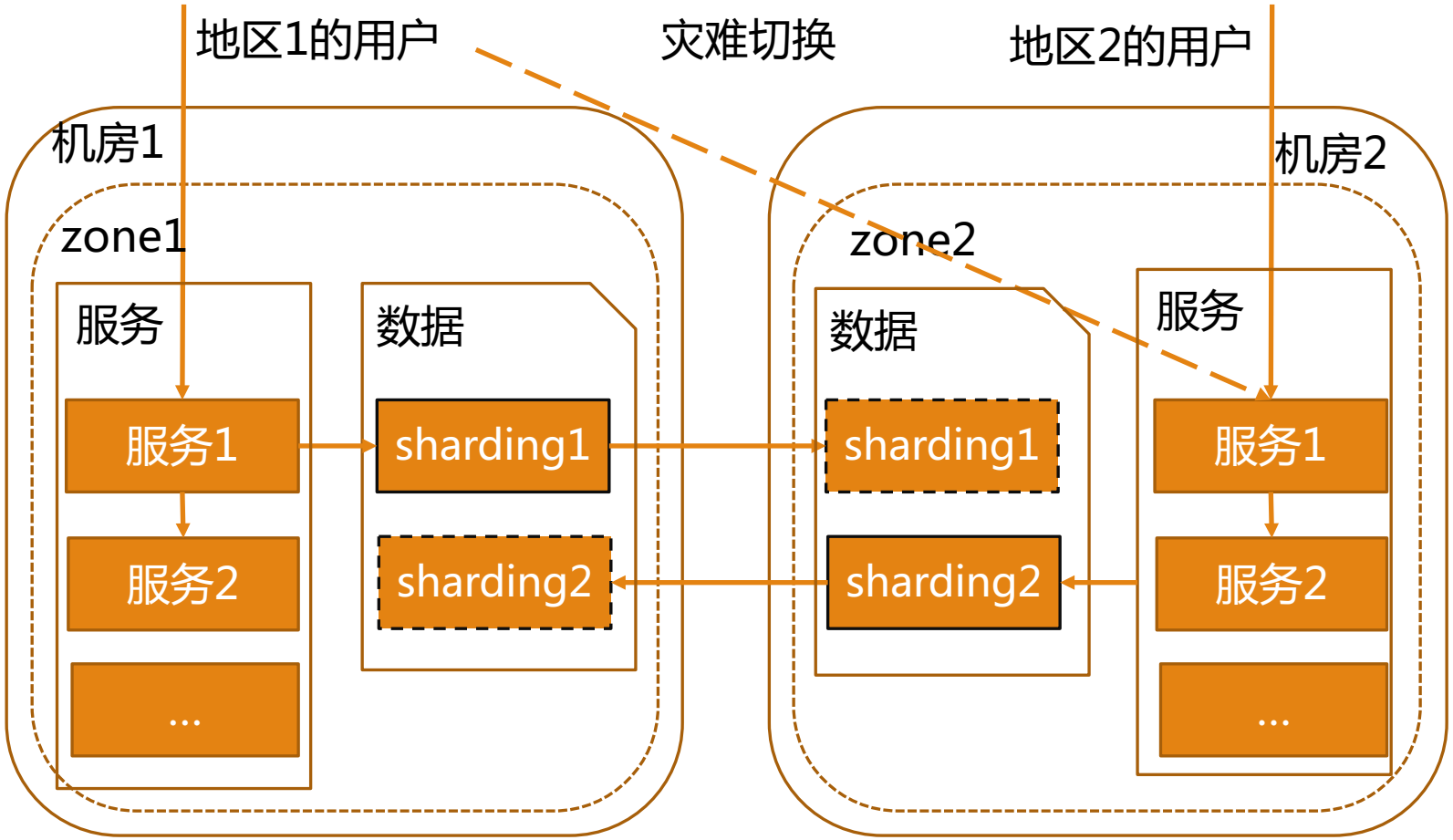
饿了么Zookeeper集群管理下一步展望

Why?


用户请求
机房1
服务1
服务2
...
机房2
服务3
服务4
...


订单量
Millions
10
9
8
7
6
5
4
3
2
1
0
2013
2014
2015
2016.5
2016.12

多活


地区1的用户
灾难切换
地区2的用户
机房1
机房2
zone1
zone2
服务
数据
数据
服务
服务1
sharding1
sharding1
服务1
服务2
sharding2
sharding2
服务2
...
...

作用

饿了么很多基础中间件都将数据保存在Zookeeper中：

- 微服务框架
- 数据库访问中间件
- 负载均衡中间件

Agenda

饿了么异地数据中心多活

- Zookeeper跨数据中心的方案

饿了么Zookeeper跨数据中心方案

饿了么Zookeeper集群管理下一步展望

What is Zookeeper?

核心特性

Reliable and consistent data store

使用的场景

- configuration information
- naming, group services
- distributed synchronization (distributed lock, leader election)

同类产品比较--Etcd, Consul

	Zookeeper	Etcd	Consul
configuration information	✓	✓	✓
naming, group services	✓	✓	✓
distributed synchronization	✓	✓	✓
Data store	✓	✓	✓
开发语言	Java	Go	Go
Reliable and consistent	Zab	Raft	Raft
Features	临时节点 (Ephemeral) 变更通知 (Watches)	租约 变更通知	Session 变更通知
	顺序节点(Sequence)	Revision 多键条件事务	Index CAS（check-and-set） Acquire lock

Zookeeper的跨数据中心方案

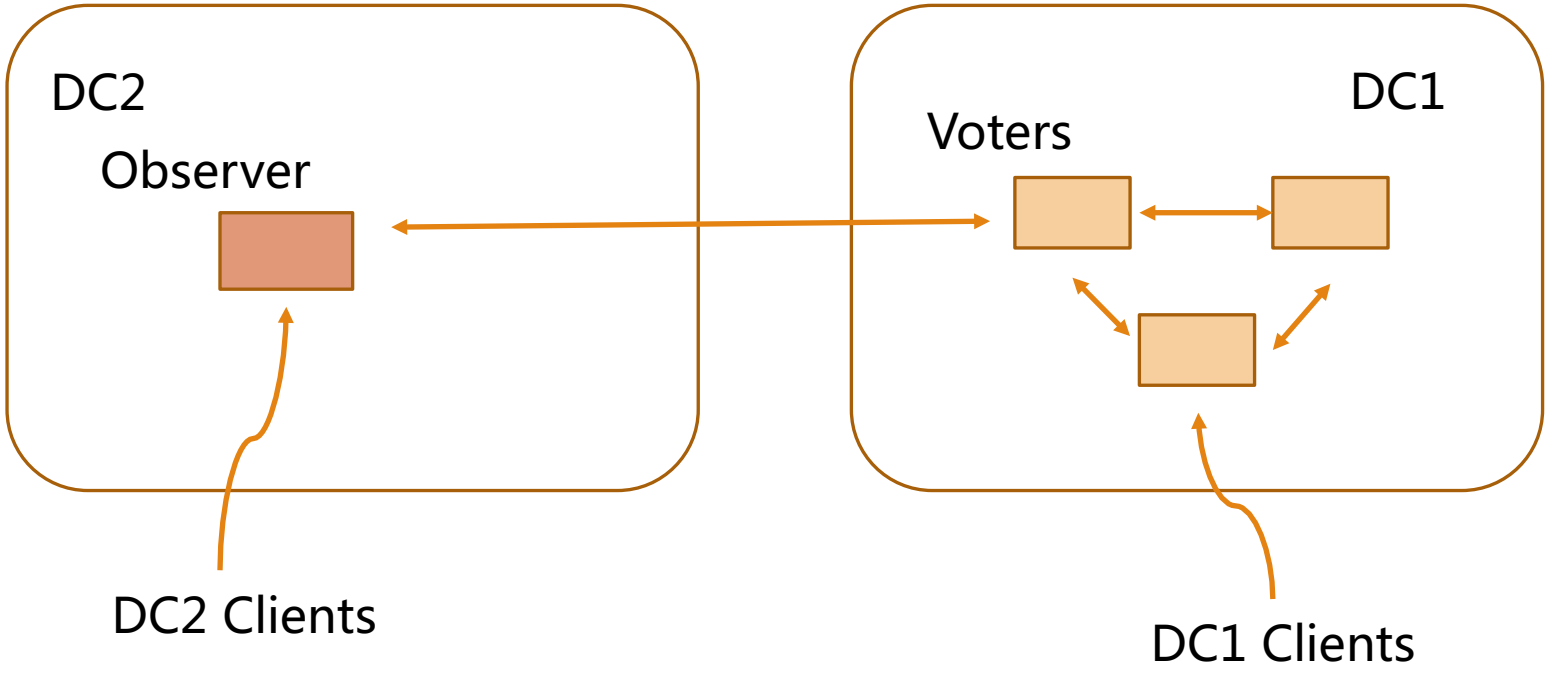
3种Zookeeper跨数据中心的方案：

Observer 方式

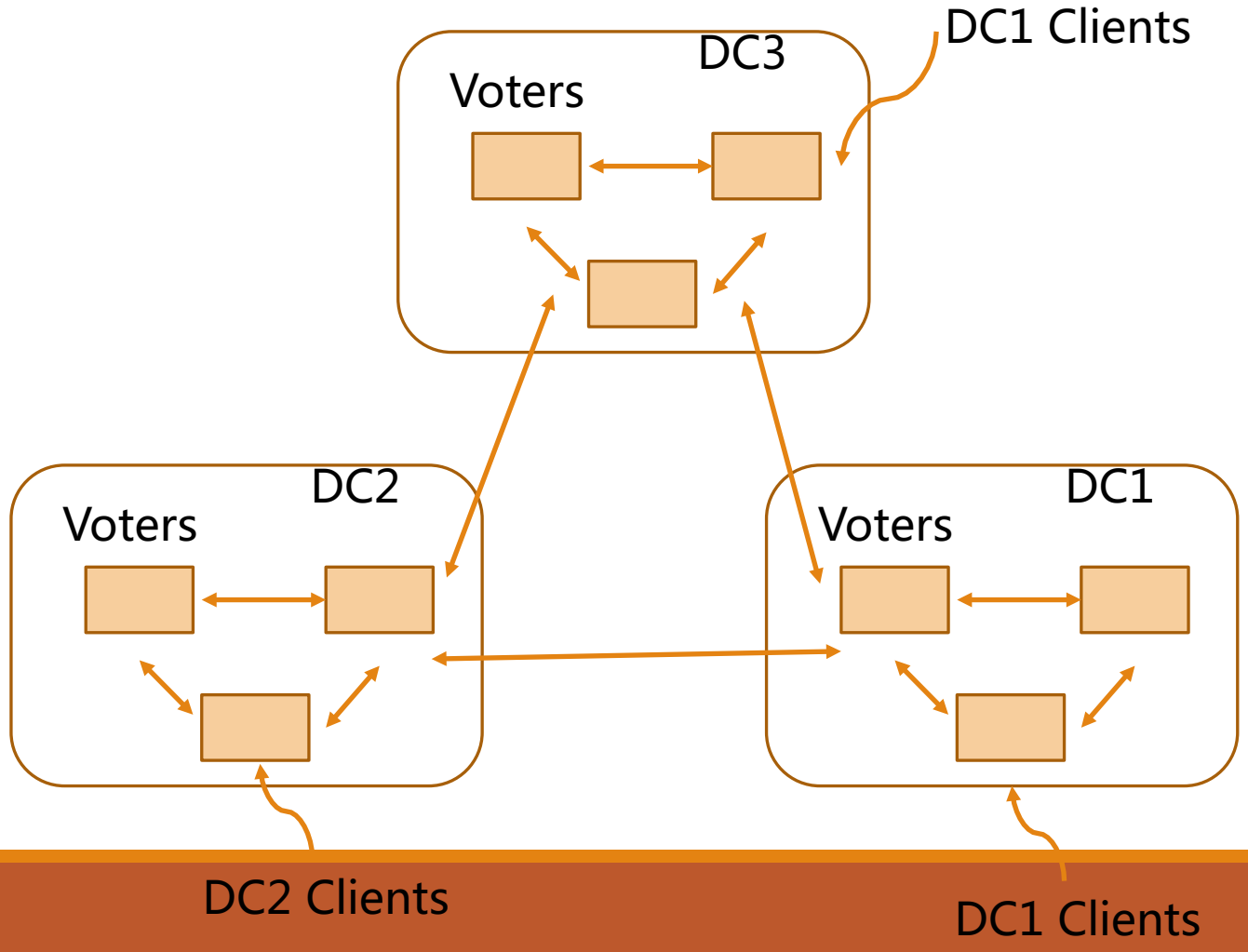
统一集群方式

独立集群方式

方案1：Observer 方式



方案2：统一集群方式


DC1 Clients
DC3
Voters
DC2
Voters
DC1
Voters
DC2 Clients
DC1 Clients

方案2：统一集群方式

统一集群方式有多种不同的部署方式：

（9节点）3：3：3

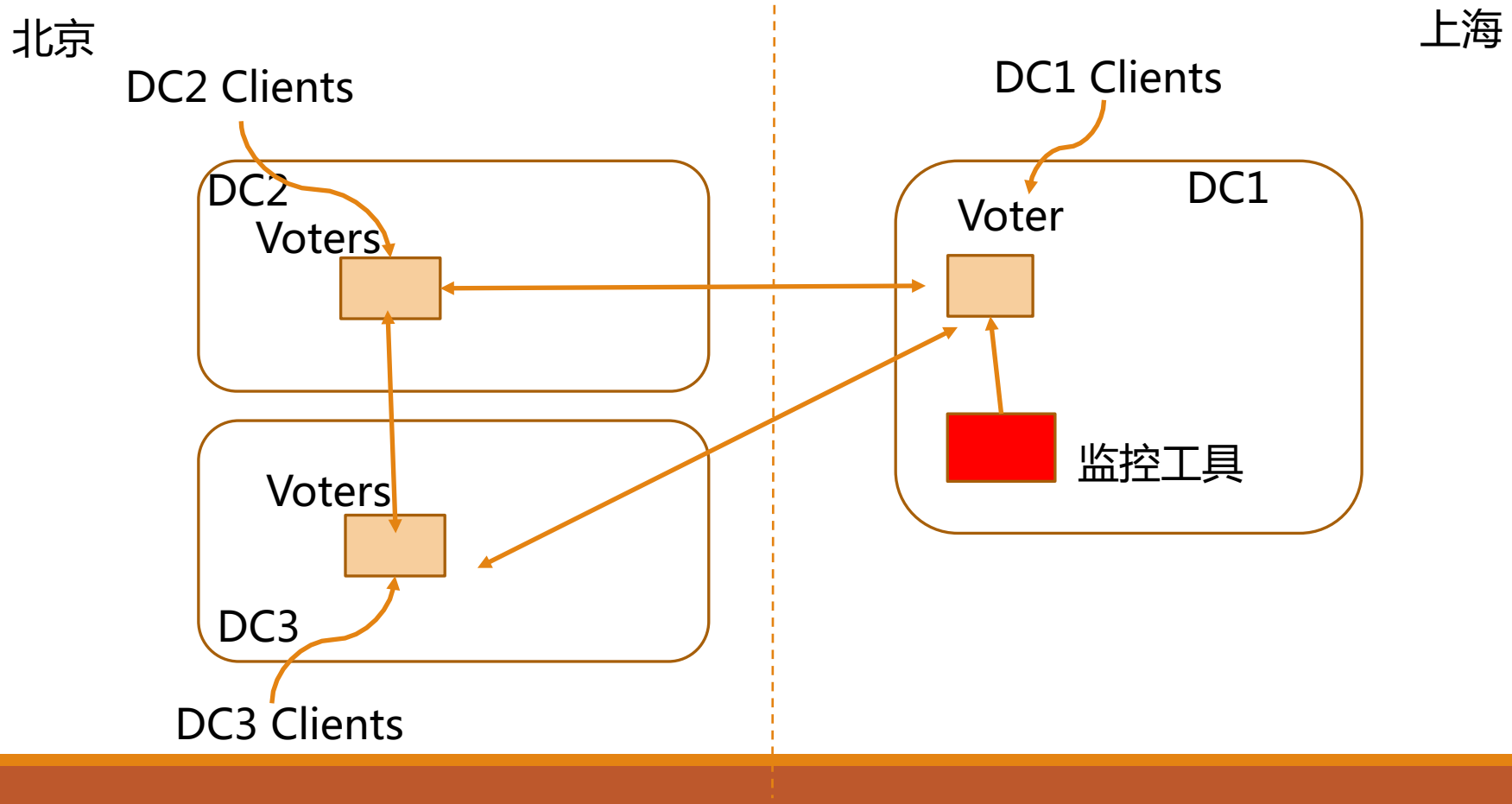
（7节点）3：2：2

（7节点）3：3：1

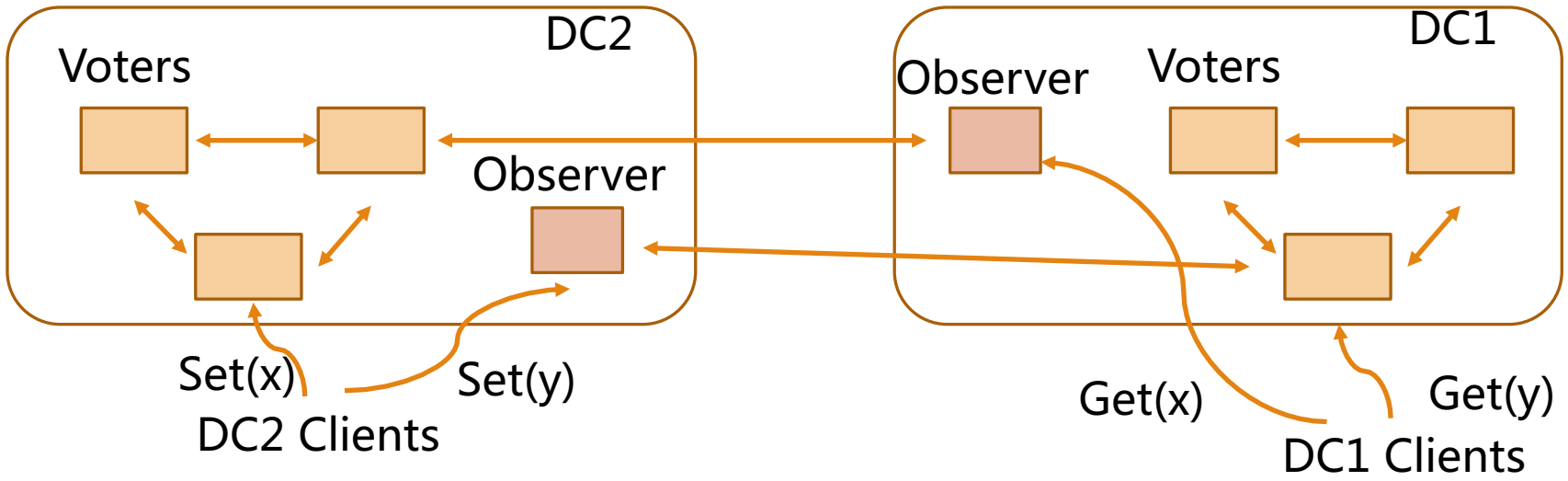
（5节点）2：2：1

等等...

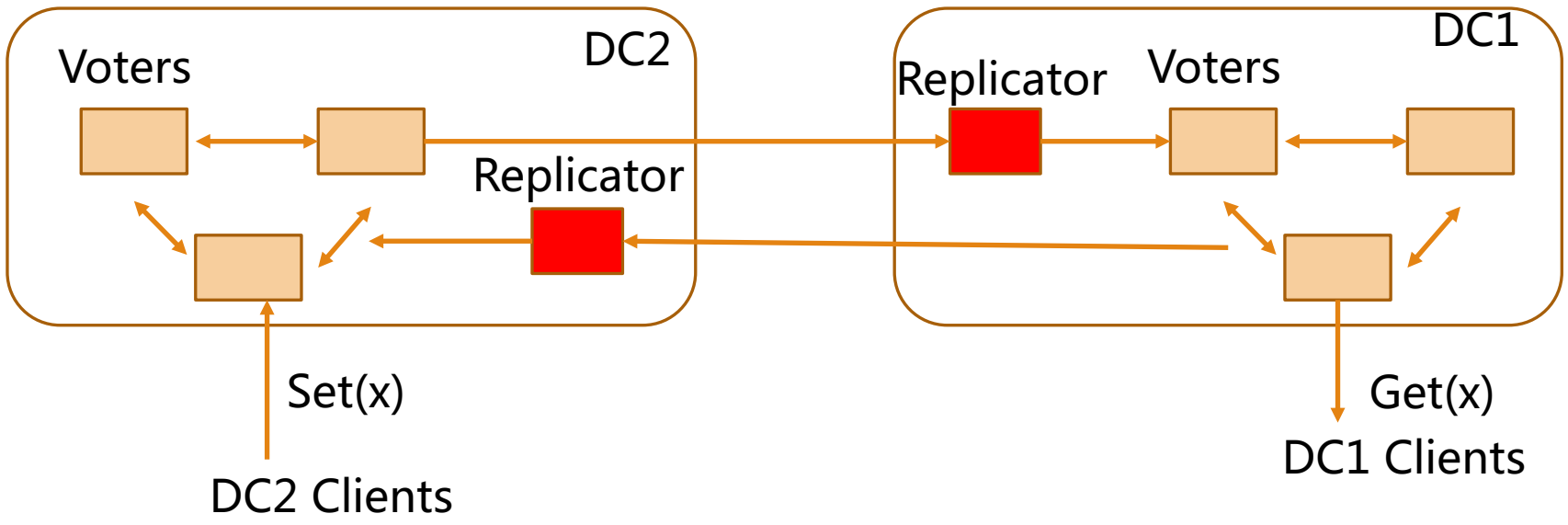
方案2：统一集群方式


北京
上海
DC2 Clients
DC1 Clients
DC2
Voters
Voter
DC1
Voters
监控工具
DC3
DC3 Clients

方案3：独立集群方案--客户端写入不同的数据



方案3：独立集群方案--数据异步复制



3种方案的总结

方案	写入性能	读性能	分区时写入	全局数据一致性
统一集群方案	非常慢	快	占大多的	是
Observer方案	慢	快	拥有voter的	是
独立集群方案	快	快	本地的	否

没有好与不好的差别，要看哪种更适合

Agenda

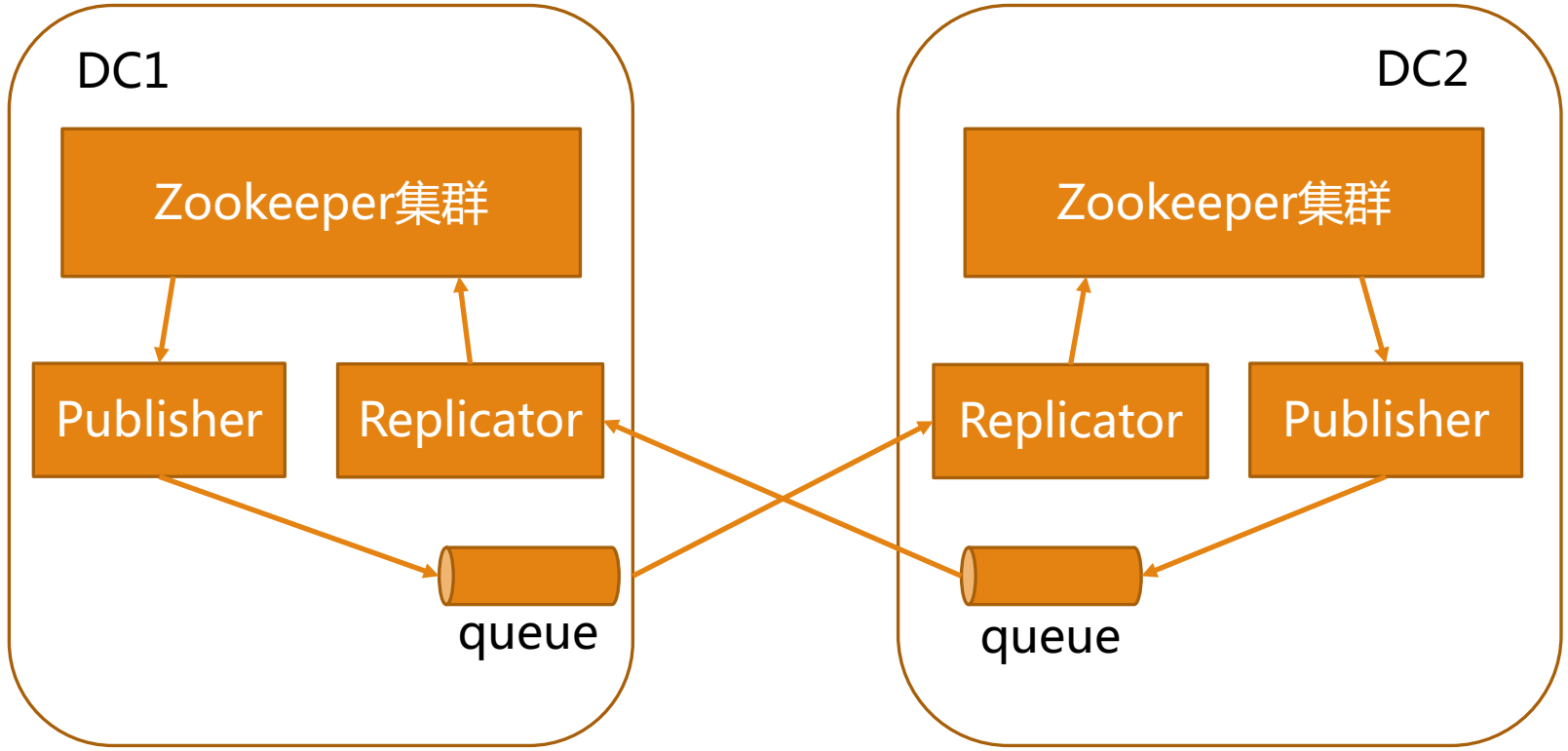
饿了么异地数据中心多活

Zookeeper跨数据中心的方案

- 饿了么Zookeeper跨数据中心方案

饿了么Zookeeper集群管理下一步展望

我们的架构


DC1
Zookeeper集群
Publisher
Replicator
queue
DC2
Zookeeper集群
Replicator
Publisher
queue

原因

多活项目更期望可用性而不是性能和一致性

存在异地机房，跨机房专线网络抖动

使用方改动尽量小

不要求支持跨机房的分布式锁

使用方可以避免数据冲突

使用方容忍少量的延时存在

Agenda

饿了么异地数据中心多活

Zookeeper跨数据中心的方案

饿了么Zookeeper跨数据中心方案

- 饿了么Zookeeper集群管理展望

What is the next?

拆分集群

数据复制中间件的改进

Zookeeper作为基础设施服务化

谢谢大家！